Catherine Cyrot Consultante

Le 12 juin 2009

Réingénierie de thesaurus : une étude de cas

L'objet de cet article est de faire partager l'expérience de réingénierie du thesaurus de l'Agence Française de Développement (AFD)[1], menée au cours de l'année 2008. Il nous a semblé que les analyses faites en préalable à ce travail de réingénierie et les conclusions qui en avaient été tirées pouvaient aider à la réflexion de professionnels d'autres secteurs placés devant ces mêmes problématiques, par exemple :

- ✓ Quels critères de choix entre la création d'un thesaurus et la réutilisation d'un logiciel existant ?
- ✓ Comment utiliser le thesaurus comme passerelle pour échanger entre organismes partenaires du même domaine d'activité ?
- ✓ Quelle utilisation au delà de l'indexation des références bibliographiques pour structurer des informations dans la perspective d'une gestion de contenu ?
- ✓ Comment situer le thesaurus dans la famille des référentiels et faire évoluer son utilisation vers des outils du web sémantique ?

Sur ces quelques questions et d'autres, il nous a semblé utile de faire partager nos analyses et les méthodologies utilisées. Elles sont abordées à la lecture de cet article, mais les lecteurs pourront la compléter en lisant sur le site www.xxx.xx une version plus complète de l'article. Enfin dans le cadre du partenariat établi avec la Bibliothèque Dag Hammarskjöld des Nations Unies à New York, une version anglophone de cet article devrait prochainement être publiée en ligne.

1. Critères de choix du thesaurus

On reprendra ici la demande de l'Agence Française de Développement (AFD) pour en dégager des perspectives plus générales. Ce sera l'occasion de mentionner les éléments qui président aux décisions de réingénierie d'un thesaurus.

1.1. La demande de l'agence française de développement

La prestation demandée par l'AFD a pour but de mettre en place un thesaurus pour indexer ses bases de données documentaires. Ce thesaurus a vocation à terme à être intégré dans les applications documentaires actuellement gérées par le logiciel CINDOC.

Ce thesaurus devait donc répondre aux demandes suivantes :

⇒ Correspondre aux **champs thématiques** de la documentation de l'AFD dans le domaine de la coopération internationale, et pour cela analyser les nomenclatures utilisées par les documentalistes.

[1] Pour mieux connaître l'AFD, on peut consulter son site Web à l'adresse http://www.afd.fr

⇒ Proposer une structure de thesaurus compatible avec les **normes internationales**, pour faciliter les échanges de l'AFD avec son réseau de partenaires, en particulier l'Institut de Développement du Sussex, et les partenaires de l'EADI[2]. Le multilinguisme serait un atout important dans cette perspective.

⇒ Etre proposé dans un **format** qui puisse être intégré dans l'application documentaire utilisée, ici celle de la société CINDOC. Ce logiciel ayant fait le choix du format XML comme format d'import export, le choix de ce standard peut permettre ultérieurement de travailler avec d'autres plateformes. Cet aspect sera développé ultérieurement à propos de la méthodologie du projet (voir § 2).

⇒ Etre réalisé avec un bon **rapport coût/efficacité**. Cette réalisation implique une double compétence documentaire et informatique. Elle nécessitera des développements informatiques étroitement associés à l'analyse des contenus du thesaurus. Cette interaction entre étude documentaire et développements informatiques se fera de manière moins structurée et plus collaborative que par le passé en privilégiant la souplesse et la réactivité. Elle suit en cela l'évolution de la gestion des projets suite à l'arrivée du Web 2[3].

L'analyse de ces exigences peut être utilisée comme une grille de lecture pour un professionnel qui situé dans la même perspective de réingénierie de thesaurus, serait confronté aux mêmes questions. En effet, ces éléments : champs thématique, conformité avec les normes internationales, problèmes de format, coût/efficacité, sont bien ceux qui président aux décisions de réingénierie d'un thesaurus.

1.2. L'analyse de la demande

La demande de l'AFD partait bien du besoin d'indexer les bases de données documentaires. Mais le traitement de ces données documentaires ne représente qu'une part des produits et services qu'elle propose à ses agents. Par ailleurs l'AFD a en projet un intranet pour la gestion de contenu de ses informations. Fallait-il s'en tenir à la commande initiale ou prévoir que ce thesaurus pourrait dans une étape ultérieure **être utilisé pour l'ensemble des contenus gérés par l'AFD**. C'est la seconde proposition qui a été retenue. On rejoint là une évolution plus générale des référentiels ; élaborés par les documentalistes, ils peuvent être utilisés bien au-delà de leur objectif premier : indexer et rechercher des données bibliographiques.

RESITUER LE THESAURUS DANS LA FAMILLE DES REFERENTIELS

L'équipe projets a resitué le thesaurus dans **la famille des référentiels** : classifications, ontologies de sorte que le thesaurus en chantier puisse évoluer vers une base de connaissance et servir de pont entre des outils complémentaires.

✓ On peut prévoir de le faire évoluer vers une ontologie en spécifiant les types de relations qu'il met en œuvre. Comme le précise Bruno Bachimont[4], l'ontologie est bien « *la signature fonctionnelle et relationnelle, munie de sa sémantique, d'un langage formel de représentation et manipulation de connaissances* », ce dernier précise ainsi les limites du thesaurus liées à la sous-détermination des relations structurant des termes, « *les relations hiérarchiques ne proposent pas …une sémantique formelle contraignant de manière stricte la relation hiérarchique,… la relation d'association n'est pas autrement spécifiée et peut renvoyer à toute sorte de relation sémantique entre les termes* ».

---

[2] Association Européenne de recherche et de Formation en matière de développement : http://www.eadi.org/

[3] BALMISSE, Gilles, OUNI, Ali. « *Gérer autrement les projets* ». Documentaliste, 2009, vol 46, n°1 p. 44-46

[4] BACHIMONT, Bruno. *Ingénierie des connaissances et des contenus : le numérique entre ontologies et documents*. Paris : Hermès, 2007. p. 233

Il en conclut « *La règle est donc simple même si elle est souvent oubliée : si l'utilisateur des index est humain, il vaut mieux prendre un thesaurus, si l'utilisateur doit être une machine comme le supposent les travaux autour du web sémantique, les ontologies s'imposent* ». Ce travail complémentaire devrait permettre d'ajouter plus de sémantique au thesaurus, en augmentant et en précisant les rapports entre les concepts. Il s'agit ainsi de revoir les concepts ou classes qui définissent un objet, et pour chaque objet ses attributs (le terme retenu est bien un attribut du concept), les relations taxinomiques (hiérarchisation des concepts) ou sémantiques qui rapprochent deux objets. Il nous semble que le fait d'articuler plus étroitement concepts et relations permettrait à l'AFD de mieux décrire le contenu de ses documents. Cette évolution est indissociable de celle d'interfaces qui intègrent les outils techniques du web sémantique.

✓ En extraire un plan de classement peut se révéler particulièrement utile dans le cadre d'un portail pour accéder aux ressources documentaires de l'AFD. Ces ponts entre thesaurus et taxinomie sont par exemple proposés par l'outil Mondeca qui permet de créer une taxinomie par extraction des niveaux du thesaurus. En cela la structure du thesaurus UNBIS conçue à partir de micro-thesaurus divisés en chapitres ou sous-domaines[5] se révèle particulièrement pertinente.

L'AFD souhaitait que ce thesaurus soit un des outils au service du développement de partenariats. Cette demande mettait en avant des problèmes de compatibilité avec les bases de données des partenaires de l'AFD. Et donc une **recherche des thesaurus existants** et utilisés dans le domaine de la coopération internationale.

Fallait-il alors :

⇒ Créer un thesaurus sur mesure, ce qui pouvait correspondre à la demande initiale ?

⇒ Retenir un thesaurus existant pour indexer les bases de données documentaires de l'AFD ?

1.3. L'analyse de l'existant

L'étude des thesaurus existant dans l'environnement des Nations Unies nous a semblé un préalable indispensable pour éclairer ces choix. On rejoint bien ici la démarche préconisée, notamment dans la norme NF Z47-101 et dans la littérature professionnelle sur la construction des thesaurus.

1.3.1. Critères à retenir pour une analyse de l'existant

Pour éclairer la question qui précède, entre construction d'un thesaurus et réutilisation d'un thesaurus existant, on a tenté de dégager des critères. Ces derniers, en phase avec la demande préalablement décrite, pourraient être utilisés plus globalement en amont d'un travail de réingénierie.

⇒ **Adaptation au domaine d'activité.** Le domaine d'activité de l'AFD était il un domaine particulier, nécessitant de ce fait un vocabulaire spécifique ? Pour répondre à cette question on a analysé le nombre de termes utilisés pour chacun de ses domaines. Cette lecture des thématiques retenues par les documentalistes montrait bien que globalement on y retrouvait les champs d'intervention des Nations Unies. Cette analyse était confortée par le volume d'informations directement issues des Nations Unies dans le référentiel bibliographique.

[5] Dans cet article, on privilégiera le terme chapitre par rapport au terme sous domaine pour décrire l'organisation des micro-thesaurus en sous-parties.

Thème AFD	Nombre de termes
Administration	26
Agriculture élevage pêche	37
Aide humanitaire	11
Collectivités Outre-mer	16
Commerce	28
Coopération internationale	25
Démographie	14
Droit	153
Economie	61
Education Formation	64
Energie	17
Entreprise	37
Environnement	20
Finances	56
Santé	126
Travail	26

Répartition des descripteurs par micro-thesaurus à l'AFD

Le vocabulaire utilisé renvoyait bien aux caractéristiques d'un macro-thesaurus[6], plutôt qu'à un vocabulaire spécialisé. Cette analyse plus fine des domaines a donc permis d'écarter les thesaurus établis dans un domaine spécifique pour retenir ceux de type macro-thesaurus.

⇒ **Multilinguisme** : la volonté d'avoir un vocabulaire multilingue était une contrainte forte pour un organisme engagé dans la coopération internationale en lien étroit avec le système des Nations Unies. Sa demande portait principalement sur un vocabulaire pivot entre français, anglais, espagnol. Prendre en compte cette demande dans la création d'un thesaurus alourdissait de manière notable le travail à réaliser et sortait du rapport coût/efficacité recherché. Le choix d'un thesaurus déjà multilingue était donc un avantage significatif. Dans son principe, ce multilinguisme fonctionne par relation d'équivalence linguistique des termes dans les différentes langues avec le terme du langage source (ici l'anglais). Chaque langue proposant des synonymes écartés qui lui sont propres.

⇒ **Maintenance** du thesaurus : ce sont pour une bonne part les problèmes de maintenance qui ont provoqué la demande de l'AFD. Le thesaurus initialement utilisé était celui d'IBISCUS, déjà compatible avec le macro-thesaurus des Nations Unies, mais qui n'avait pas été remis à jour depuis 1998. Dans l'intervalle, les documentalistes l'avaient enrichi avec des candidats descripteurs sans en contrôler la structure d'ensemble. Au vu de cette expérience, le choix d'un thesaurus dont la maintenance soit efficace était un critère de choix important.

[6] « *Thesaurus couvrant un très large domaine et ne comprenant que des descripteurs d'un niveau général qui peuvent être repris dans des descripteurs spécialisés* » (cf. Norme NF Z47-100)

1.3.2. Un tableau comparatif des thesaurus disponibles dans l'environnement des Nations Unies

L'équipe de projet a dressé en préalable à ses choix un tableau comparatif rapide des thesaurus disponibles par recherche des thesaurus existant[7]. Les partenaires de l'AFD se situent principalement dans la mouvance des Nations Unies, ce qui restreignait la recherche. Pour les principaux thesaurus analysés, les critères suivants ont permis de classer les solutions préférentielles. : Editeur, domaine d'activité, multilinguisme, maintenance, nombre de descripteurs.

Nom	éditeur	domaine	multilinguisme	maintenance	Nombre de descripteurs
UNBIS	Bibliothèque Dag Hammarskjöld des Nations Unies	Programmes et activités de l'Organisation des Nations Unies	Anglais, arabe, chinois, espagnol, français russe	Révision en continu	7033 descripteurs
UNESCO	UNESCO	Education, culture, sciences sociales et humaines	Anglais, français, espagnol, russe	La dernière version date de 2008	7000 descripteurs
AGROVOC	FAO	Agriculture, pêche, alimentation, environnement	Anglais, français, espagnol, arabe, chinois, …	Dernière mise à jour 2008	30 000 descripteurs
Thesaurus du BIT	BIT	Travail	Anglais français espagnol	Dernière édition en 2008	4000 descripteurs
Thesaurus du commerce international	CNUCED/OMC	Commerce international	Anglais ,espagnol	Edition 2004	?
Thesaurus de la Banque Mondiale	Banque mondiale	Economie internationale	Anglais	Révision en continu	?

Sélection de thesaurus des Nations unies, source thesaurus UNBIS

Il faut noter que le thesaurus UNBIS fait figure de macro-thesaurus compatible avec les autres thésaurus spécialisés des agences des Nations Unies qui font autorité dans leur domaine d'activité respectif[8]. C'est à ce titre que les choix de descripteurs dans ces domaines spécialisés ont été privilégiés par UNBIS.

[7] Pour voir la liste des thesaurus disponibles dans le système des Nations Unies : http://lib-thesaurus.un.org

[8] La liste de ces thesaurus peut être consultés sur le site UNBIS

Notre proposition a donc été de réutiliser un de ces thesaurus existants en l'adaptant aux besoins de l'AFD. Nous détaillerons ultérieurement les ajustements faits pour répondre à sa demande.

1.4. En conclusion

C'est en croisant ces trois critères, adaptation au champ thématique, multilinguisme et qualité de la maintenance que le choix de l'AFD s'est porté sur le thesaurus UNBIS. Elle a donc validé la décision de retenir le thesaurus UNBIS comme le plus proche de ses domaines d'intervention. Le choix du multilinguisme s'est porté sur l'anglais et l'espagnol. Le fait que le thesaurus UNBIS soit accessible en ligne et géré par une équipe de documentalistes avec des mises à jour régulières a été un élément déterminant de choix.

2. La méthodologie de fabrication du thesaurus

Une fois le choix du thesaurus UNBIS adopté, il restait à travailler la compatibilité de ce vocabulaire avec la collection de termes utilisés par l'AFD. Il ne nous semblait pas possible de contraindre les indexeurs à utiliser tels quels les descripteurs du thesaurus UNBIS. Une large part correspondait à l'usage de la communauté des usagers, mais pas tous. Certains descripteurs parfois traduits directement de l'anglais semblaient incongrus dans l'environnement de travail de l'AFD. Enfin, des concepts spécifiques à l'environnement professionnel en France ne se retrouvaient pas dans le thesaurus source. Il ne s'agissait pas de traduire en français le thesaurus UNBIS, puisqu'une des langues de travail de ce thesaurus était le français, on était plutôt devant un travail de compatibilité[9] entre le thesaurus UNBIS considéré ici comme langage source et le vocabulaire de l'AFD considéré comme vocabulaire cible. Il s'agissait bien alors de faire un travail de mise en relations entre termes pour établir des correspondances entre les termes utilisés par l'AFD et les descripteurs du thesaurus source par analyse de leurs équivalences :

- ✓ Ajout de nouveaux descripteurs,
- ✓ Descripteurs changés en non descripteurs,
- ✓ Non descripteurs changés en descripteurs,

Pour garantir la compatibilité ultérieure avec le thesaurus UNBIS, les modifications portant sur les descripteurs par rapport au thesaurus UNBIS, ont été communiqués aux gestionnaires de ce thesaurus dans une logique de rétroaction vers le vocabulaire source. Par contre, les nombreux ajouts de synonymes ou quasi-synonymes introduits dans les relations d'équivalence qui ne modifiaient pas la structure du thesaurus source ont été réalisés en local.

On a cherché à résoudre les principaux problèmes posés par la fusion de vocabulaires tels que les décrit Sylvie Dalbin[10] :

- ✓ Structure hiérarchique : La restructuration des hiérarchies a été évitée autant que possible, et la notion de terme de tête[11] (top term) a été conservée.
- ✓ Adaptation de la sémantique ; la proximité et l'habitude de travailler dans un environnement international limitait les glissements de sens entre thesaurus source et thesaurus cible.

---

[9] Rappelons que selon la norme NF Z46-100, « *deux thesaurus A et B sont compatibles si dans un même domaine les documents indexés avec le thesaurus A peuvent être retrouvés avec le thesaurus B et réciproquement* ».

[10] DALBIN, Sylvie. « Thésaurus et informatique documentaire : partenaires de toujours ». *Documentaliste*. 2007, vol. 44, n° 1, p. 42-55

[11] Terme de tête = top term c'est-à-dire le terme situé au plus haut de la hiérarchie ascendante du descripteur.

On a pu relever les cas de figure suivants :

- Equivalence inexacte,
- Equivalence partielle,
- Equivalence d'un terme à plusieurs termes,
- Spécificité et usabilité du vocabulaire obtenu non satisfaisante.

Quand l'équipe projet a rencontré ces cas de figure, elle a travaillé pour repérer dans le thesaurus source UNBIS le concept caché derrière le terme utilisé en remontant au langage pivot utilisé (ici l'anglais) et en explorant son environnement sémantique. En complément, elle a mieux défini son territoire en rédigeant des notes d'application à partir de l'étude de la terminologie existante.

Elle a alors recherché le concept le plus proche dans les termes couramment utilisés par l'AFD et analysé les relations d'équivalence (synonymie, synonymie proche, variantes lexicales, renvoi au générique) afin de rapprocher les cas de figure dans lesquels des glissements sémantiques pouvaient s'opérer entre les termes UNBIS et ceux de l'AFD. Le fait de pouvoir générer des relations d'équivalence différentes selon les langues dans la structure UNBIS a permis ces adaptations ; le plus souvent une note d'application précise l'utilisation du terme.

Exemples

Terme UNBIS	Terme retenu pour l'AFD
Accroissement de la population	Croissance démographique
Isolement social	Exclusion
Retrait de capitaux	Fuite des capitaux
Sexe social	Genre
Folklore	Culture populaire

Exemples d'adaptation sémantique

- ✓ L'adaptation de la forme : quand les termes retenus étaient au pluriel, habitude courante dans les thesaurus anglo-saxons, ils ont été conservés tels quels dans le thesaurus pour rester le plus proche possible du thesaurus source, même si cela écarte de la norme française NF Z47-100 qui recommande l'utilisation du singulier quand cela ne crée pas d'ambiguïté de sens.

La méthodologie utilisée pour compatibiliser ces deux vocabulaires est détaillée au §2.2 de cet article.

2.1. La méthodologie informatique appliquée par Actonis

Le travail a été réalisé par une équipe associant compétences documentaires et informatiques. La société Actonis dont il sera fait mention ultérieurement a géré le projet pour cette équipe.

Au niveau informatique, cinq étapes principales ont été identifiées :

1. Construction d'une base de données reflet du thesaurus UNBIS, à partir d'un fichier unique, reprenant toutes les informations du thesaurus des Nations Unies (voir aussi § 2.2.1),
2. Construction d'une base de données reflet de la base documentaire de l'AFD. (voir aussi § 2.3),
3. Un ensemble de traitements informatiques d'analyse de la base documentaire, et de rapprochement des termes utilisés par l'AFD. avec ceux du thesaurus UNBIS (voir aussi § 2.3),

4. Un ensemble de traitements informatiques pour générer le nouveau thesaurus dans l'application CINDOC (voir aussi § 2.4),
5. Une mise à jour de l'indexation du référentiel documentaire par des requêtes SQL fournies par ACTONIS (voir aussi § 4.1).

La méthodologie a étroitement associé les traitements documentaires et informatiques. Tous les traitements ont été déportés sur les ressources techniques d'ACTONIS, ce qui a permis d'avoir une très grande souplesse de développement, sans contrainte pour les utilisateurs de l'AFD. Les résultats obtenus étaient systématiquement testés avant que leur résultat soit fourni à l'AFD.

Pour réaliser ces traitements, on a en particulier utilisé :

- ✓ Une plate-forme PHP/My Sql pour importer la base de données et générer les traitements,
- ✓ Un éditeur de texte pour les manipulations de fichiers entre ces interfaces,
- ✓ Une version du logiciel Altova pour analyser et valider les formats XML,
- ✓ Une version du logiciel CINDOC pour analyser les fichiers bibliographiques fournis par l'AFD en entrée et tester en sortie les traitements réalisés.

2.2. Format du thesaurus source des Nations Unies

2.2.1. Au niveau du format des données

Un fichier unique de 698 023 lignes a été fourni par la Bibliothèque Dag Hammarskjöld des Nations Unies au format UTF-8[12]), chaque enregistrement pouvant comporter un nombre variable de caractères jusqu'à 800 caractères. Pour chaque enregistrement le libellé de l'attribut était suivi de son contenu, le nombre d'attributs étant inconnu et variable.

Concernant la première étape du travail informatique, qui consistait à créer une base de données reflétant le thesaurus UNBIS, la principale difficulté provenait de l'absence documentation technique concernant la constitution du fichier fourni par les Nations Unies.

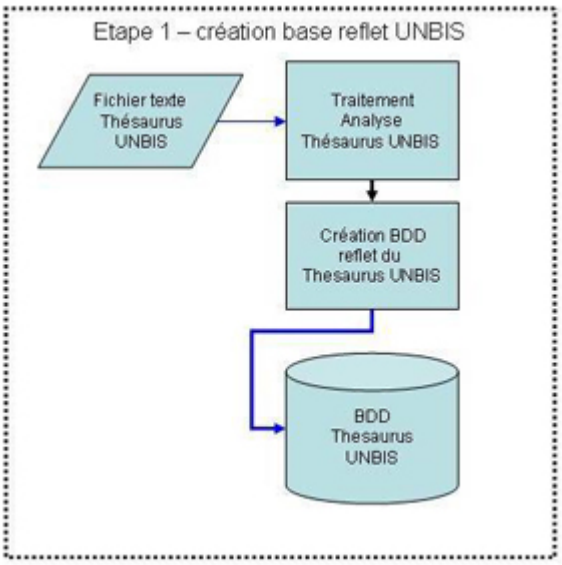
Le travail préalable a consisté à :

- ✓ Récupérer la structure du thesaurus, après l'avoir analysée pour retrouver la liste des champs utilisés, dans une première analyse, jusqu'à 104 attributs ont pu être relevés.
- ✓ Générer à partir de cette analyse **une 1ère base de données de 7033 termes** qui liste la totalité des enregistrements, après avoir filtré les caractères parasites qui en bloquaient la lecture.
- ✓ Analyser les contenus pour ne garder au final que les **champs utiles**, Un travail d'analyse a été fait pour corréler les attributs fournis par UNBIS avec la structure de thesaurus gérée par CINDOC. 23 attributs sur les 104 ont été retenus pour créer une seconde base de données SQL (en particulier les attributs correspondant à des langues non prévues au cahier des charges, ou les attributs correspondant à des informations dupliquées et éventuellement concaténées n'ont pas été retenus).

[12] UTF-8 (UCS transformation format 8 bits) est un format de codage de caractères défini pour les caractères Unicode (UCS). L'IETF (Internet engineering Task Force) requiert qu'UTF-8 soit pris en charge par les protocoles de communication d'Internet échangeant du texte (http://fr.wikipedia.org/wiki/UTF-8).

✓ Définir **l'organisation des données**, les relations entre les champs et pour chaque champs son mode de fonctionnement. On a ainsi défini les modalités de transformation de la structure récupérée dans la base de données MySQL vers le fichier XML à importer dans le thesaurus géré par CINDOC (voir § 2.3).

ETAPE INFORMATIQUE 1


Etape 1 – création base reflet UNBIS
Fichier texte Thésaurus UNBIS
Traitement Analyse Thésaurus UNBIS
Création BDD reflet du Thesaurus UNBIS
BDD Thesaurus UNBIS

Pour chaque terme du thesaurus UNBIS, un nombre variable de champ était affecté. De plus, le fichier fourni présentait les informations verticalement. Chaque champ avait lui-même une longueur indéterminée. Ceci a demandé la mise en place d'un algorithme d'analyse particulier, par approches successives, avant de déterminer et de programmer un algorithme fiable pour environ 99,9 % des cas. Pour moins d'une dizaine de termes il a fallu intervenir manuellement sur la base de données constituée.

Pour donner à l'AFD une vision plus concrète du thesaurus et expliquer ces étapes préalables, un export d'une centaine de références test a été édité. Ce thesaurus test a été commenté lors d'une réunion avec les documentalistes.

2.2.2. Au niveau conceptuel

Le thesaurus est construit en respectant les normes de construction des thesaurus monolingues (NF Z47-100) et multilingues (NF Z47-101)

RECAPITULATIF DE LA STRUCTURE MULTILINGUE DU THESAURUS

On a donc pour chaque descripteur

⇒ Sa traduction dans les différentes langues,

⇒ Sa note d'application dans les différentes langues,

⇒ Son rattachement à un chapitre et sous-chapitre du micro-thesaurus, déclinés dans les différentes langues,

⇒ Son rattachement à un terme générique dans les différentes langues et son rattachement au terme de tête (top term),

⇒ Ses termes spécifiques dans les différentes langues,

⇒ Ses termes associés dans les différentes langues,

⇒ Ses synonymes ou quasi-synonymes dans les différentes langues,

LA HIERARCHIE

Exemple
03.03.00 - ENVIRONNEMENT
Nombre de termes inclus dans le thème: 45
ACCORDS SUR L'ENVIRONNEMENT
DEGRADATION DE L'ENVIRONNEMENT
DEVELOPPEMENT DURABLE
DROIT DE L'ENVIRONNEMENT

Les descripteurs sont organisés en micro-thesaurus correspondant à un **plan de classement** de 18 micro-thesaurus répartis en 143 chapitres. Ces micro-thesaurus correspondent à des ensembles de descripteurs appartenant au même domaine, ces descripteurs sont ventilés à l'intérieur du micro-thesaurus en chapitres. Ce type d'organisation en micro-thesaurus divisés en chapitres nous semblait particulièrement intéressant dans la mesure où il permettait d'utiliser le thesaurus pour définir une organisation éditoriale, comme une taxinomie[13], particulièrement bien adaptée pour donner accès aux documents numériques à travers un portail d'information. Il facilitait donc un usage du thesaurus pour gérer l'ensemble des contenus numériques de l'AFD comme cela avait été envisagé lors de l'analyse de la problématique.

Le thesaurus UNBIS comportait des **polyhiérarchies**[14], un même descripteur pouvant appartenir à plusieurs micro-thesaurus et/ou chapitre ; dans ce cas il est rattaché à un terme de tête dans le sous-chapitre correspondant. Toutefois cette possibilité était, de fait, relativement peu utilisée. Comme l'indiquent les réflexions publiées sur les polyhiérarchies, leur utilisation rend complexe la gestion des thesaurus. Après l'avoir testé sur un mini-thesaurus d'une centaine de termes, la décision a été prise de les abandonner. Elles ont été remises à plat en privilégiant une relation hiérarchique principale correspondant au choix préférentiel retenu (selon les recommandations de la norme NF Z47-100). Les rattachements hiérarchiques non retenus ont été transformés en relations associatives pour conserver toute la richesse des relations du thesaurus.

Exemple

03.03.00 - ENVIRONNEMENT	16.03.00 - SCIENCES DE LA TERRE
Nombre de termes inclus dans le thème: 45	Nombre de termes inclus dans le thème: 52
ACCORDS SUR L'ENVIRONNEMENT	ACTIVITE SISMIQUE
ASPECTS ECOLOGIQUES	**BIOGEOCHIMIE**
ATTEINTE A L'ENVIRONNEMENT	CARTES GEOLOGIQUES
BIOGEOCHIMIE	CHIMIE DES SOLS
CHIMIE DE L'ENVIRONNEMENT	CRISTALLOGRAPHIE

[13] En informatique, le terme taxinomie (taxonomy en anglais) désigne une méthode de classification des informations dans une architecture structurée de manière évolutive. Le terme est couramment employé pour des systèmes de gestion de contenu (CMS) (Wikipedia)

[14] On rappellera que la polyhiérarchie telle qu'elle est décrite dans la norme NF Z47-100 revient à subordonner une notion à plusieurs notions en même temps.

LA STRUCTURE AVEC LES RELATIONS

On a bien extrait à travers les 23 attributs définis préalablement les relations classiques du thesaurus : équivalence (EM/EP), hiérarchie (TG/TS) et association (TA). Il convient ici de rappeler que ce maillage des relations des descripteurs permet de les resituer dans leur contexte et donc d'en préciser le sens.

⇒ **«** *Les relations hiérarchiques expriment les rapports de supériorité et subordination entre notions* [15]» Les termes génériques sont généralement repérés par le sigle **TG** . Les termes spécifiques, sont généralement repérés par le sigle **TS.**

⇒ **«** *Les relations associatives, encore appelées relations de voisinage, indiquent des analogies de signification entre les termes et sont utilisées pour recouvrir les relations entre les notions étroitement liées, mais évidemment en excluant les relations hiérarchiques et les relations d'équivalence* [15]». Ces termes connexes repérés généralement par le sigle **TA** (relation d'association : causalité, localisation, relations de nature temporelle, composition, …) permettent au chercheur de modifier progressivement son interrogation ou de l'élargir sans faire appel aux termes génériques.

⇒ Les notes d'application, spécifiques aux langues utilisées, permettent de mieux définir les concepts sous-tendus par les descripteurs retenus. On a vu en particulier qu'elles ont été utilisées et parfois enrichies pour gérer les glissements sémantiques entre descripteurs.

⇒ **«** *Les relations d'équivalence ou de substitution renvoient des divers synonymes ou quasi-synonymes d'un concept vers le descripteur unique correspondant utilisé dans le langage documentaire* [15]. Cette relation permet souvent le renvoi de termes spécifiques trop peu utilisés vers leurs termes génériques. Les termes équivalents sont généralement repérés par le sigle **EP** comme abréviation de « *Employé Pour* ». Le terme à préférer au terme Employé Pour est indiqué par le symbole **EM** ou **EMP** comme abréviation de « *Employer* ». Dans un thesaurus multilingue, les synonymes et quasi-synonymes peuvent être spécifiques aux langues utilisées.

Exemple
DEVELOPPEMENT DURABLE
03.03.00 - ENVIRONNEMENT
Employé pour:
ECODEVELOPPEMENT
Note d'application: Développement écologiquement viable
Termes génériques:
DEVELOPPEMENT ECONOMIQUE
Termes spécifiques:
ENERGIE DURABLE
Termes associés:
ECOTOURISME
EVALUATION DE L'IMPACT SUR L'ENVIRONNEMENT
PROTECTION DE L'ENVIRONNEMENT

LE MULTILINGUISME

Il est géré à partir de l'anglais comme vocabulaire pivot ; à chaque terme dans une langue correspond un équivalent dans les autres langues du thesaurus. Seules les relations d'équivalence et les notes d'application sont spécifiques à chacune des langues retenues.

[15] Norme NF Z47-100

On peut en consulter la présentation sur la version en ligne du thesaurus UNBIS[16], par exemple dans une recherche sur le développement durable.

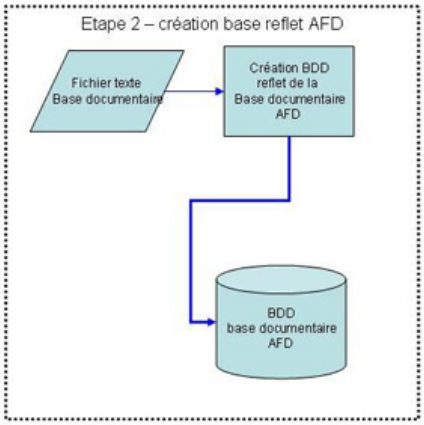
Le travail d'analyse préalable a consisté comme indiqué précédemment à ne retenir que les versions en français, anglais et espagnol des descripteurs retenus.

2.2.3. En conclusion

Le travail préalablement décrit pour récupérer le thesaurus UNBIS aurait été grandement facilité s'il avait été fourni, transcodé en SKOS[17]. Ce travail de migration du thesaurus UNBIS vers SKOS[18] serait à réaliser par l'équipe qui maintient UNBIS, comme cela a déjà été fait par exemple pour le thesaurus AGROVOC

2.3. L'analyse des micro-thesaurus utilisés par l'AFD

ETAPE INFORMATIQUE 2


Etape 2 – création base reflet AFD
Fichier texte Base documentaire
Création BDD reflet de la Base documentaire AFD
BDD base documentaire AFD

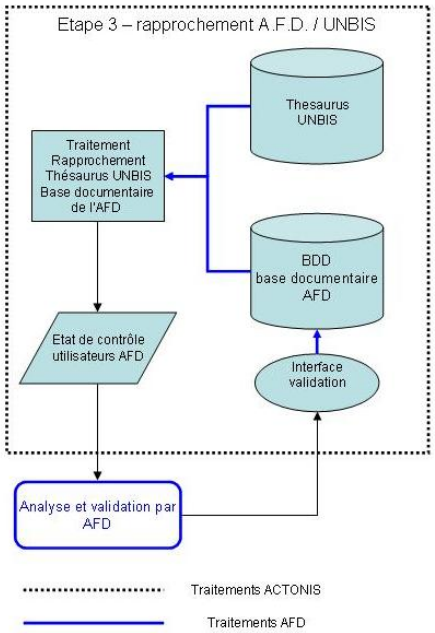
L'AFD a fourni à Actonis une fichier de 3123 mots clés utilisés pour indexer ses bases de données bibliographiques, soit :

- ✓ 1470 termes retenus,
- ✓ 1653 termes à priori écartés de la sélection.

Concernant la deuxième étape du travail informatique, la construction de la base de données reflet de la base documentaire de l'AFD n'a pas posé de difficulté, puisqu' Actonis se retrouvait dans une configuration classique de création de bases de données à partir d'éléments clairement documentés.

ETAPE INFORMATIQUE 3

Les traitements informatiques, de la troisième étape ont été créés pour rapprocher efficacement les descripteurs UNBIS et les termes utilisés dans le thesaurus de l'AFD (et donc dans la base documentaire comme mot-clé). Ils ont demandé la mise en place d'un algorithme syntaxique afin de permettre un rapprochement probable entre des termes qui étaient par exemple, au pluriel dans un référentiel et au singulier dans l'autre, notamment pour les termes composés. Ce rapprochement syntaxique a permis d'associer les termes utilisés par l'AFD à ceux du thesaurus UNBIS dans plus de 60 % des cas.


Etape 3 – rapprochement A.F.D. / UNBIS
Thesaurus UNBIS
Traitement Rapprochement Thésaurus UNBIS Base documentaire de l'AFD
BDD base documentaire AFD
Etat de contrôle utilisateurs AFD
Interface validation
Analyse et validation par AFD
Traitements ACTONIS
Traitements AFD

[16] http://lib-thesaurus.un.org

[17] SKOS ou Simple Knowledge Organisation System (Système simple d'organisation des connaissances) est une famille de langages formels permettant une représentation standard des thésaurus, construit sur la base du langage RDF par le W3C. Il permet la publication d'un thesaurus pour son utilisation dans le cadre du Web sémantique et une transition possible vers des référentiels utilisant le modèle RDF, celui des ontologies OWL par exemple (Wikipedia).

Chacun de ces 1470 termes utilisés par l'AFD, ont été analysés et comparés avec les descripteurs ou les termes équivalents du thésaurus UNBIS afin d'établir des concordances entre les termes AFD et les descripteurs UNBIS. Les descripteurs UNBIS retenus ont été validés avec les documentalistes de l'AFD lors de réunions de travail périodiques, à partir de l'analyse préalable réalisée par Actonis. Les cas de figure moins consensuels ont été revus et les choix tranchés avec la responsable de la documentation de l'AFD. Ce travail en commun avec les documentalistes de l'AFD a facilité ultérieurement l'appropriation par l'AFD du thésaurus proposé par Actonis.

- ✓ Equivalence exacte : 537 (451+86) termes soient 30% étaient identiques entre les descripteurs AFD et UNBIS validant ainsi à posteriori les analyses qui avaient précédé le choix de ce thesaurus. Ce travail de comparaison a été fait de manière entièrement automatique.
- ✓ Equivalence partielle ou inexacte : Pour 790 termes, soit 53%, on pouvait trouver un terme du thesaurus UNBIS auxquels les rattacher par une relation d'équivalence. Une première comparaison a été faite de manière automatique pour les termes du thesaurus de l'AFD correspondant strictement aux termes écartés du thesaurus UNBIS. Dans un second temps une analyse des racines des termes de l'AFD a été préparée par un traitement automatique puis validée. Elle a permis de retrouver les variantes morphologiques des descripteurs, par exemple ceux utilisés au pluriel dans le monde anglo-saxon et au singulier dans le monde francophone. Ce travail de comparaison a été complété par une analyse intellectuelle sur les termes suffisamment proches pour générer des relations d'équivalence.

  Pour 66 termes, des équivalences ont été trouvées, mais pour des raisons culturelles, ce sont les termes de l'AFD qui ont été retenus, couramment utilisés par le public francophone de préférence aux termes du thesaurus UNBIS qui en étaient des traductions de l'anglais. *Exemple : prestation de service pour services de support.* On se trouvait bien là devant des glissements sémantiques pour lesquels une analyse spécifique a été réalisée.
- ✓ Non-Equivalence : 30 descripteurs jugés indispensables pour indexer des documents francophones ont été créés pour être rattachés ultérieurement à l'arborescence UNBIS. Ils correspondaient à des concepts propres à la Coopération française, n'existant pas dans le thesaurus UNBIS, *Exemple : coopération décentralisée.*

  Enfin, 67 termes, retenus dans un 1er temps, n'ont pas été pris en compte pour le thesaurus, jugés in fine trop spécifiques, ils pourront être réutilisés comme candidats descripteurs.

Une comparaison automatique a ensuite été faite entre les 1653 descripteurs écartés par l'AFD et les descripteurs UNBIS, leurs synonymes ou quasi-synonymes[19] écartés par analyse des relations d'équivalence d'UNBIS. L'analyse intellectuelle de cette comparaison a abouti à réintégrer 133 termes dans la liste des descripteurs retenus, soit comme descripteurs, soit comme équivalences.

Le même travail d'analyse, plus simple a été réalisé pour les 260 descripteurs géographiques.

[19] ». La norme NF Z47-100 définit les quasi-synonymes comme « *des termes dont la signification est généralement différente dans l'usage normal mais que l'on peut considérer comme synonymes pour les besoins particuliers de l'indexation* »

2.4. L'import dans le module thesaurus de CINDOC

Ce travail a consisté à exporter les descripteurs et leurs relations de la base de données SQL au format XML. Il faut noter que le format XML retenu par CINDOC n'est pas compatible avec SKOS. Réutiliser le thesaurus hors du champ de l'application documentaire comme souhaitaient le faire les commanditaires de l'AFD nécessiterait de le faire évoluer vers les technologies Web, en utilisant la norme SKOS. Ce qui manque pour l'essentiel au thesaurus élaboré dans ce projet pour être transcodé en SKOS est de spécifier pour chaque concept retenu et son descripteur une URI (Uniform Resource Identifier[20]).

Une étude préalable de toutes les relations figurant dans le thesaurus UNBIS a été faite et une maquette réalisée pour présenter leur transcription au format XML requis par l'interface CINDOC. Pour chacun de ces cas de figure, une note précisant les règles de fonctionnement a été rédigée avec sa traduction en XML. Ceci vaut notamment pour :

- ✓ La gestion du nom de domaine,
- ✓ La gestion des micro-thesaurus,
- ✓ L'entrée du descripteur avec son rattachement au domaine et au micro-thesaurus,
- ✓ La gestion des relations d'équivalence avec plusieurs renvois sur le terme retenu,
- ✓ La gestion des termes associés,
- ✓ La gestion des termes génériques et spécifiques,
- ✓ L'insertion des notes d'application,
- ✓ La gestion du multilinguisme en anglais et en espagnol,
- ✓ La gestion des équivalents historiques.

Cette analyse est disponible à l'adresse http://www.catherine-cyrot.net/afd/

Par exemple pour les descripteurs

- ✓ Le descripteur français (**FTerm**) est unique et obligatoire et composé éventuellement de plusieurs termes,
- ✓ Il est rattaché à un micro-thesaurus et à un chapitre à l'intérieur du micro-thesaurus,
- ✓ Il peut comporter les relations hiérarchiques, associatives et les relations d'équivalence, ces relations étant réciproques. Toutes les relations réciproques sont à déclarer de manière biunivoque dans le fichier XML.

Entrée du descripteur avec micro-thesaurus et domaine

```
<Object>
    <thesaurusEntry lang=«F»>
        <term type=«DE» status=«1»>Bovins</term>
        <domain>04 Agriculture</domain>
        <microThesaurus>04 03 00 Cheptel</microThesaurus>
    </thesaurusEntry>
</Object>
```

Gestion du terme générique

```
<Object>
    <thesaurusEntry lang=« F »>
        <term type=«DE» status=«1»>Agrométéorologie</term>
        <domain>16 Sciences et techniques</domain>
        <microThesaurus>16 07 00 Météorologie </microThesaurus>
        <topTerm>Météorologie</topTerm>
        <broaderTerm>Météorologie</broaderTerm>
    </thesaurusEntry>
</Object>
```

Gestion des synonymes (Employé employer pour)

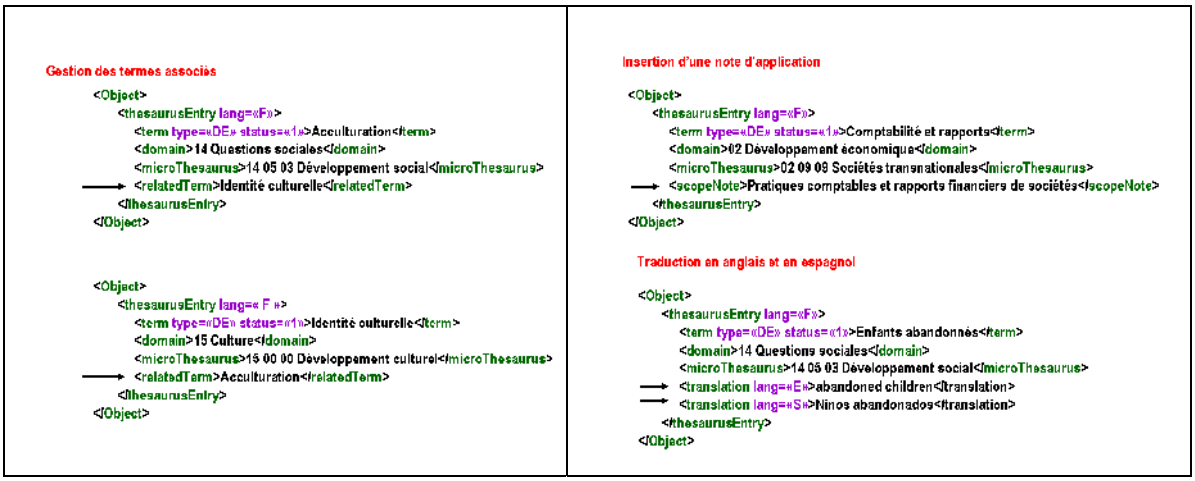
```
<Object>
    <thesaurusEntry lang=«F»>
        <term type=« DE » status=«1»>Abaca</term>
        <domain>04 Agriculture</domain>
        <microThesaurus>04 02 02 Cultures</microThesaurus>
        <usedFor>Chanvre de Manille</usedFor>
    </thesaurusEntry>
</Object>

<Object>
    <thesaurusEntry lang=«F»>
        <term type=«ND»> Chanvre de Manille</term>
        <domain>04 Agriculture</domain>
        <microThesaurus>04 02 02 Cultures</microThesaurus>
        <useInstead>Abaca</useInstead>
    </thesaurusEntry>
</Object>
```

Descripteur retenu → Employé pour

Employer →

[20] L'Uniform Resource Identifier (URI), chaîne de caractères unique respectant la norme RFC 3990 du W3C



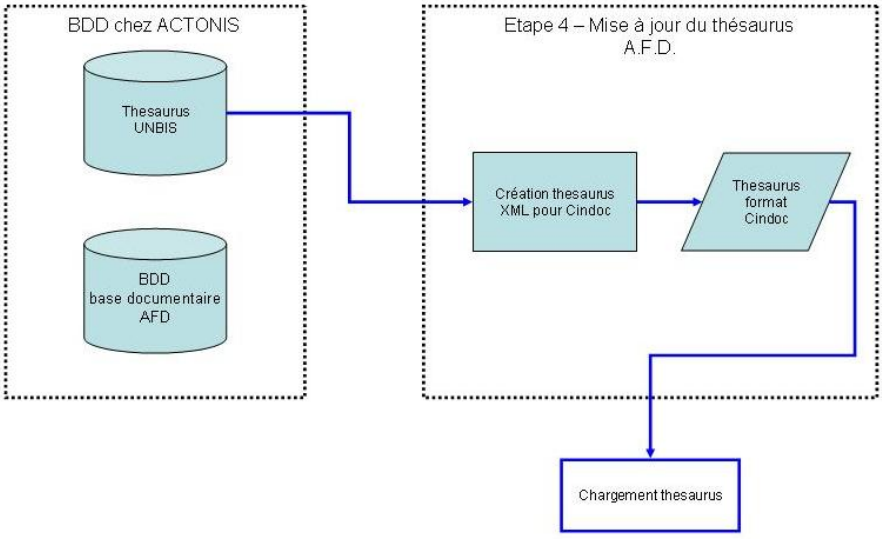
Exemple d'analyse préalable à la génération du format XML

Le choix d'XML comme format d'import dans l'applicatif CINDOC est particulièrement intéressant dans la mesure ou ce standard permet au thesaurus d'être réutilisé par d'autres applicatifs basés sur le format XML. Utiliser la norme SKOS pour le décrire améliorerait encore le standard XML retenu.

AU FINAL : ETAPE INFORMATIQUE 4

Enfin, dans la dernière étape, la génération d'un fichier XML pour créer le nouveau thesaurus de l'AFD a nécessité un travail particulièrement pointu, dans la mesure où la validation du nouveau thesaurus par l'applicatif CinDoc, ne donnait pas d'éléments exhaustifs en cas de rejet du fichier XML lors de la validation des relations hiérarchiques entre les termes.

Etape 4 Mise à jour des référentiels


BDD chez ACTONIS
Thesaurus UNBIS
BDD base documentaire AFD
Etape 4 – Mise à jour du thésaurus A.F.D.
Création thesaurus XML pour Cindoc
Thesaurus format Cindoc
Chargement thesaurus

Le fichier xml du thesaurus une fois l'import dans le gestionnaire de thesaurus testé en local sur les machines d'Actonis a été fourni à l'AFD avec un diaporama détaillant la procédure d'import. : procédure qui a été réalisée avec succès.

On présentera ici un exemple d'utilisation du gestionnaire CINDOC

3. L'organisation du thesaurus

3.1. La structure du thesaurus résultat

La structure du thésaurus utilisée par l'AFD a été gardée le plus proche possible de celles fournie par UNBIS. Le thesaurus comporte 1307 descripteurs. Sur ces 1307 descripteurs, 834, soit environ 2/3 sont des syntagmes[21] et 473, soit environ 1/3 sont des unitermes. On a donc plutôt privilégié le choix de syntagmes, ils permettent de localiser une information très précise, regrouper des termes étroitement associés, décrire des concepts complexes. On a privilégié la précoordination[22] de termes pour définir des notions complexes correspondant à des expressions fréquemment utilisées, plutôt que de laisser à l'utilisateur le soin de post-coordonner les descripteurs en les combinant avec des opérateurs booléens, ce qui risque d'augmenter le bruit dans sa recherche. Ces choix traduisent une évolution significative par rapport à la norme dont les préconisations sur ce point nous paraissent obsolètes. Plus proches du langage naturel, les syntagmes semblaient plus en phase avec l'évolution des langages et leur utilisation à venir par des outils du web sémantique, on rejoint là les analyses de Muriel Amar[23].

ORGANISATION EN MICROTHESAURUS

Ces descripteurs sont répartis en 17 micro-thesaurus et 125 chapitres. Chaque micro-thesaurus est donc structuré en 7 chapitres en moyenne. L'importance de cette structuration explique pour une part que les hiérarchies soient relativement courtes comme expliqué ci-dessous. Elle devrait permettre une double utilisation possible du thesaurus : navigation dans la structure

[21] Un syntagme est un ensemble de mots formant une seule unité catégorielle et fonctionnelle, mais dont chaque constituant, parce que dissociable (contrairement au mot composé), conserve sa signification et sa syntaxe propres. Un syntagme constitue donc une association occasionnelle, libre, alors que le mot composé est une association permanente (lorsqu'un syntagme se fige, il devient bien sûr un composé détaché, soit une locution). (Wikipedia)

[22] « *Principe suivant lequel l'expression d'une notion s'effectue au cours de l'élaboration du thesaurus par la création de termes composés* » (cf norme NF Z47-100)

[23] AMAR, Muriel. *Les fondements théoriques de l'indexation : Une approche linguistique*. Paris : ADBS, 1997. 355 p.

des relations du thesaurus et utilisation d'un plan de classement, par exemple pour l'afficher sur le portail pour accéder à tous les documents posés sur l'intranet.

RELATIONS HIERARCHIQUES

Le fait que les hiérarchies soient courtes, comme on peut le voir dans l'analyse des niveaux de hiérarchie du thesaurus retenu a simplifié le travail de structuration du thesaurus. Ce choix est cohérent avec celui d'un vocabulaire riche en syntagmes plus difficiles à intégrer dans des structures en arbre. Le thesaurus de l'AFD suit en cela l'évolution générale des thesaurus vers des grappes hiérarchiques plus courtes.

Nombre de relations hiérarchiques	529
Hiérarchies de 1er niveau	356
Hiérarchies de second niveau	130
Hiérarchies de troisième niveau	33
Hiérarchies de quatrième niveau	10

On a gardé la notion de **terme de tête** (top term), prise en compte dans les normes américaines et anglaises mais non dans les normes plus anciennes de l'AFNOR et de l'ISO. Par contre, suivant en cela la pratique du thesaurus UNBIS, on n'a pas spécifié le type de hiérarchie : hiérarchie d'instance de type « *est un spécimen ou est un exemple d'une catégorie* » ou hiérarchie partitive de type « *tout/partie* » qui ont fait leur apparition dans les normes américaine et anglaise. Mieux spécifier ces relations permettrait de passer du thesaurus à une base de connaissances s'appuyant sur des ontologies.

Les 30 termes nouveaux ont été rattachés aux branches des hiérarchies UNBIS, de la manière la plus logique possible.

RELATIONS ASSOCIATIVES

Par contre, 1135 termes ont des relations associatives, soit 87% des descripteurs. C'est bien l'enrichissement des relations associatives qui a été privilégié dans la construction de ce thesaurus par rapport à la profondeur hiérarchique. Cette option est intéressante dans la mesure ou ces relations associatives pourraient être retravaillées comme des ontologies, si on les qualifie ultérieurement, ce qui constituerait un outil précieux pour une application de type web sémantique. La encore ces relations sont présentes dans les normes américaines et anglaises, leur prise ne compte permettrait de passer d'une structure de type thesaurus à une structure de type base de connaissance.

RELATIONS D'EQUIVALENCE

L'enrichissement du thesaurus initial a été pour l'essentiel constitué de l'apport des termes écartés réintroduits dans une relation d'équivalence de type EM/EP (employé/employer pour). Quand la relation d'équivalence a été inversée, le terme retenu par l'AFD a pris la place du terme UNBIS dans la hiérarchie du thesaurus. Au total, 1713 non descripteurs renvoient aux 1307 descripteurs retenus. Augmenter le nombre de non-descripteurs revient à multiplier les points d'entrée dans le thesaurus.

Ce travail d'enrichissement permettra dans un premier temps de faire évoluer l'indexation des bases de données bibliographiques, il serait utile pour enrichir ultérieurement les dictionnaires utilisés par des logiciels de Web sémantique. La encore, cette évolution vers la multiplication des points d'entrée est une tendance assez générale de l'évolution des thesaurus.

EN CONCLUSION

On voit bien que si le thesaurus de l'AFD intègre des évolutions récentes, qui permettraient une meilleure utilisation par les outils du web sémantique, cette évolution s'arrête encore en chemin. En faire un référentiel normatif adapté pour gérer des bases de connaissances supposerait un travail complémentaire pour mieux en spécifier les relations, et intégrer en sus des éléments demandés par les définitions de type de document (DTD) du langage XML [24] (date de création ou de mise à jour, source du terme, noms des gestionnaires, lien avec d'autres langages, …)

3.2. Une édition du thesaurus imprimé au format PDF

L'édition réalisée propose, en accord avec les normes existantes, une présentation alphabétique des descripteurs avec toutes leurs relations, une présentation hiérarchique et une liste d'unitermes. Cette édition constitue un outil de travail pour les documentalistes de l'AFD sous ses 3 formes :

- ✓ **Liste alphabétique des descripteurs** avec toutes leurs relations : Outre les relations d'équivalence de hiérarchie et d'association, elle intègre le code de classification et le terme de tête (MV) ainsi que les notes d'application et les traductions en anglais (E) et espagnol (S), reprenant en cela les préconisations de la norme anglaise. En outre intégrer les termes de tête (ou top terms) contourne en partie le fait que cette présentation ne donne qu'une vision partielle de la hiérarchie entre les termes quand, comme c'est le cas ici, un seul niveau de hiérarchie est indiqué.

COMMERCE EXTERIEUR
[E] FOREIGN TRADE
[S] COMERCIO EXTERIOR
[DO] 07 - COMMERCE INTERNATIONAL
[MT] 07.01.00 - COMMERCE INTERNATIONAL ET POLITIQUE COMMERCIALE
[MV] COMMERCE INTERNATIONAL
[TG] COMMERCE INTERNATIONAL
[TS] EXPORTATIONS
[TS] IMPORTATIONS
[VA] BALANCE COMMERCIALE
[VA] COMMERCE INTERIEUR
[NE] Du point de vue d'un pays ou d'une région du monde déterminée

RELATIONS COMMERCIALES
[EM] COMMERCE EXTERIEUR

Exemple de liste alphabétique de descripteurs

[24] La Document Type Definition (DTD), ou Définition de Type de Document, est un document permettant de décrire un modèle de document SGML ou XML. (Wikipedia)

- ✓ **Liste hiérarchique**, elle reprend la présentation par domaine d'intérêt.

```
[DO] 01 - QUESTIONS POLITIQUES ET JURIDIQUES
 [MT] 01.01.00 - INSTITUTIONS ET MOUVEMENTS POLITIQUES
        ASPECTS POLITIQUES
        AUTONOMISATION
        COLLECTIVITE TERRITORIALE
              ADMINISTRATION LOCALE
                    ADMINISTRATION MUNICIPALE
              COOPERATION DECENTRALISEE
        COMMUNES
```

Exemple de liste hiérarchique de descripteurs

- ✓ **Liste permutée** : le fait que le thesaurus comporte un grand nombre de descripteurs multi-termes renforce l'intérêt de cette présentation en présentant dans un ordre alphabétique les mots significatifs qui composent les descripteurs. La liste permutée intègre les descripteurs et non descripteurs, comme l'indique la norme anglaise (la norme américaine ne les inclut pas et la norme française ne précise rien).

CONTRAT	ADMINISTRATIF	[EM] CONTRATS
DROIT	ADMINISTRATIF	
ETABLISSEMENT PUBLIC	ADMINISTRATIF	[EM] ETABLISSEMENTS PUBLICS
SERVICES DE SUPPORT	ADMINISTRATIF	[EM] PRESTATIONS DE SERVICE
ACTES	ADMINISTRATIFS	
FRAIS	ADMINISTRATIFS	
	ADMINISTRATION	
	ADMINISTRATION DECENTRALISEE	[EM] DECENTRALISATION
	ADMINISTRATION DOUANIERE	

Exemple de liste permutée de descripteurs

3.3. La maintenance du thesaurus

La maintenance du thesaurus a été prévue dès sa conception.

LE SUIVI DE SON UTILISATION

Il a été prévu de réaliser après six mois de son utilisation, une analyse des termes utilisés et des candidats descripteurs :

Dans le cadre d'une formation à l'indexation pour les documentalistes, elle a permis de préciser les règles d'utilisation des descripteurs : sélectivité des concepts et profondeur d'indexation, modalités de création de candidats descripteurs, …

- ✓ Pour préparer le mode d'emploi après avoir analysé les candidats descripteurs susceptibles d'être, selon les cas :

Transformés en descripteurs, quand ce sont des notions nouvelles ou des termes plus spécifiques dont l'utilisation devient courante dans l'activité de l'AFD,

Introduits comme relations d'équivalences quand ce sont des synonymes, quasi-synonymes ou variantes lexicales de descripteurs existants,

Laissés comme candidats descripteurs quand leur usage ne justifie pas de les introduire dans le thesaurus,

Corrigés quand ce sont des erreurs (fautes d'orthographe, …).

UN MODE D'EMPLOI POUR SA MAINTENANCE

Un guide a été réalisé et fourni à l'AFD pour définir les modalités d'évolution du thesaurus en le gardant compatible avec le thesaurus UNBIS. Ce travail de mise à jour s'appuie sur une évaluation critique périodique (environ deux fois par an) des descripteurs utilisés pour l'indexation.

Dans tous les cas, la décision du groupe de travail est préparée par l'administrateur du thesaurus à partir des candidats descripteurs introduits par les documentalistes dans un champ autonome ;

- ✓ Il fait une analyse sémantique du concept sous-jacent sur les sites qui constituent des référents terminologiques pour le domaine[25] et auprès des spécialistes de l'AFD pour en préciser le sens,
- ✓ Il lance la comparaison des candidats introduits avec le nouveau thesaurus de l'AFD pour contrôler que ces derniers n'existent pas déjà dans le thésaurus comme descripteurs ou comme équivalents,
- ✓ Il recherche dans le thesaurus UNBIS si le descripteur n'existe pas déjà ou si le candidat peut être introduit dans une relation d'équivalence avec un descripteur existant ; dans ce cas, il retient de préférence le terme UNBIS pour enrichir le thesaurus de l'AFD.

Ce travail de mise à jour s'appuie sur les contrôles automatiques des relations sémantiques du gestionnaire de thesaurus CINDOC, il vérifie en particulier l'unicité des termes, la réciprocité des relations d'équivalence, de hiérarchie et d'association, la présence de termes orphelins.

[25] Par exemple la base de données terminologique des Nations Unies, http://unterm.un.org/

4. L'interaction avec l'application documentaire de l'AFD

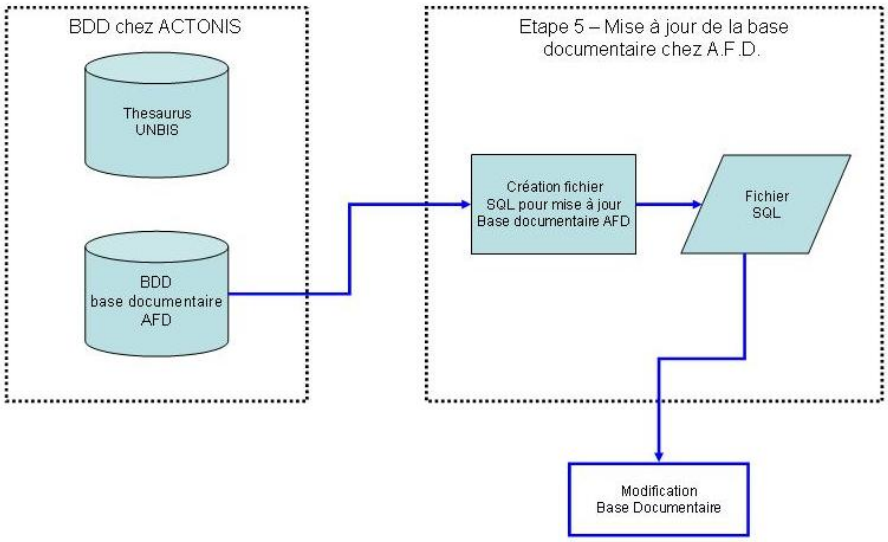
4.1. La remise à niveau de l'indexation de la base bibliographique

En décembre 2007, la base de données bibliographique comportait 12 531 références, 94% des références ayant 2802 descripteurs et seulement 200 références des candidats descripteurs, cette possibilité étant donc au total peu utilisée. Le nombre moyen de descripteurs par référence était de 4,43. L'indexation matière était introduite dans un champ MOTSCLES pour les termes de l'ancien thesaurus, mais on a vu précédemment que les documentalistes s'étaient peu à peu écartés des listes existantes pour déterminer de nouveaux descripteurs, aucun contrôle n'étant fait en saisie sur ce champs. Un champ CANDIDAT était prévu pour une indexation libre.

A partir du moment où les documentalistes utilisent un nouveau thesaurus se posait la question de la remise à niveau des indexations existantes dans les fichiers bibliographiques. Un module a donc été mis en place pour mettre à niveau cette indexation. Pour préparer la remise à niveau des fichiers bibliographiques, Actonis a récupéré la base de données bibliographique en utilisant une procédure d'import/export.

Le travail a consisté à relire le champ MOTCLES avec une requête SQL sur la base de données et dans un premier temps à le sauvegarder. Les termes d'indexation ont été comparés systématiquement avec les descripteurs et reversés dans le champ MOTSCLES s'ils étaient identiques. S'ils étaient identiques à des termes écartés, le descripteur pointé par la relation d'équivalence était reversé dans le champ MOTSCLES. Tous les termes non trouvés dans le thesaurus étaient conservés en CANDIDAT. Au final, on a donc un champ MOTSCLES avec les descripteurs qui appartiennent au thesaurus (anciens termes d'indexation ou descripteurs traduits à partir de la relation d'équivalence employer/employé pour) et un champ candidat avec les descripteurs qui n'appartiennent pas au thesaurus. Ce champ candidat continue à être utilisé par les documentalistes comme outil pour faire évoluer ultérieurement le thesaurus (voir § 3.3).

Dans cette étape, la génération du fichier permettant la mise à jour de la base documentaire de l'AFD ne présentant aucune difficulté puisqu'il s'agissait d'utiliser un langage standard, à savoir le langage SQL


BDD chez ACTONIS
Thesaurus UNBIS
BDD base documentaire AFD
Etape 5 – Mise à jour de la base documentaire chez A.F.D.
Création fichier SQL pour mise à jour Base documentaire AFD
Fichier SQL
Modification Base Documentaire

4.2. L'utilisation du thesaurus

Depuis l'utilisation du thesaurus, en juillet 2008, soit sur 9 mois, 954 références nouvelles ont été créées, et 649 descripteurs, soit environ 50% des descripteurs ont été utilisés.

A L'HEURE ACTUELLE

Le thesaurus a été branché sur la base de données bibliographique ; il est utilisé en saisie pour visualiser les descripteurs et leurs relations, et en contrôler la saisie.

Il permet des recherches guidées dans la base de données bibliographique en visualisant les descripteurs et leurs relations avec la possibilité de naviguer dans les relations du thesaurus et de sélectionner les descripteurs retenus avec éventuellement une fonction d'autopostage[26].

DANS UN AVENIR PROCHE

L'utilisation du thesaurus pourrait être étendue à l'ensemble des produits documentaires réalisés par l'AFD, et permettre un accès guidé aux archives de ces produits documentaires :

- ✓ Le **Kaléidoscope** qui propose tous les mois un package avec un ouvrage récent + 1 article + un rapport Web pour une thématique retenue, un pays, un débat sur l'aide au développement, une rubrique « *à ne pas manquer* ».
- ✓ **Produitdoc,** trimestriel accessible dans les publications de l'AFD pour suivre les cours des matières premières.
- ✓ Les **bibliographies** réalisées à la demande pour les utilisateurs de l'AFD.

DANS UN AVENIR PLUS LOINTAIN

Le thesaurus pourrait être utilisé pour des recherches sur l'ensemble des documents posés sur le site intranet de l'AFD, documents qui ne sont pas à ce jour facilement accessibles, sauf à utiliser l'explorateur Windows et à bien connaître l'emplacement de ce que recherchent les agents de l'AFD. Cela supposerait que l'AFD se dote d'un logiciel de gestion de contenu et intègre le thesaurus dans une recherche sémantique.

L'AFD pourrait à partir du travail réalisé sur la structure du thesaurus le faire évoluer vers une base de connaissances, comme cela a été mentionné à plusieurs reprises dans cet article. Cette évolution suppose, d'une part, qu'à chaque descripteur corresponde une URI stable. D'autre part, cela suppose de qualifier de manière plus précise le type de relations, qu'elles soient hiérarchiques ou associatives.

A partir du thesaurus, l'AFD pourrait extraite une taxinomie, afin de disposer d'un plan de classement pour faciliter la navigation dans un portail documentaire à reconstruire dans le but de faciliter l'accès des agents aux documents déposés sur leur Intranet.

Ces pistes d'évolution ont été évoquées avec les documentalistes lors d'une journée de formation pour les aider à mieux utiliser le nouveau thesaurus et définir ses modalités de maintenance. Plusieurs applications de ce type, actuellement disponibles sur le Web, ont été utilisées pour explorer les possibles utilisations futures du thesaurus.

[26] Expansion automatique de la recherche aux relations du descripteurs : prise en compte des relations hiérarchiques, des relations associatives, des relations d'équivalence.

5. En conclusion,

On récapitulera rapidement

Les problèmes rencontrés, ils provenaient pour l'essentiel :

- ✓ De fichiers sources insuffisamment documentés, ou spécifiques par rapport aux normes couramment utilisées (la norme SKOS par exemple),
- ✓ De procédures informatiques insuffisamment maîtrisées au niveau applicatif.

Les clés de succès

- ✓ Une méthodologie rigoureuse pour traiter les relations du thesaurus, dans le respect de normes reconnues,
- ✓ Une bonne interaction entre travail documentaire et développements informatiques au sein de l'équipe Actonis, et un test rigoureux des procédures sur la plateforme Actonis à toutes les étapes du traitement,
- ✓ Un travail d'analyse et de suivi qui impliquait étroitement à toutes les étapes la participation des documentalistes de l'AFD, ce partenariat étroit a facilité l'appropriation de ce nouvel outil de travail.

Bibliographie

AMAR, Muriel. *Les fondements théoriques de l'indexation : une approche linguistique*. Paris : ADBS, 1997. 355 p.

BACHIMONT, Bruno. *Ingénierie des connaissances et des contenus : le numérique entre ontologies et documents*. Paris : Hermès, 2007. 279 p.

MANIEZ, Jacques. *Actualité des langages documentaires : fondements théoriques de la recherche d'information*. Paris : ADBS, 1994. 395 p

« Référentiels, données d'autorité, thésaurus, ontologies, taxonomies… Pour en savoir plus ! » *Documentaliste*. 2008, vol 45, n° 3, p. 14-15

« Langages documentaires et outils linguistiques : principes, usages, perspectives - Première partie - La représentation des contenus ». *Documentaliste*. 2007, vol. 44, n° 1, p. 17-64

DALBIN, Sylvie. Descripteurs. http://www.descripteurs.net/

Thesaurus UNBIS. http://lib-thesaurus.un.org